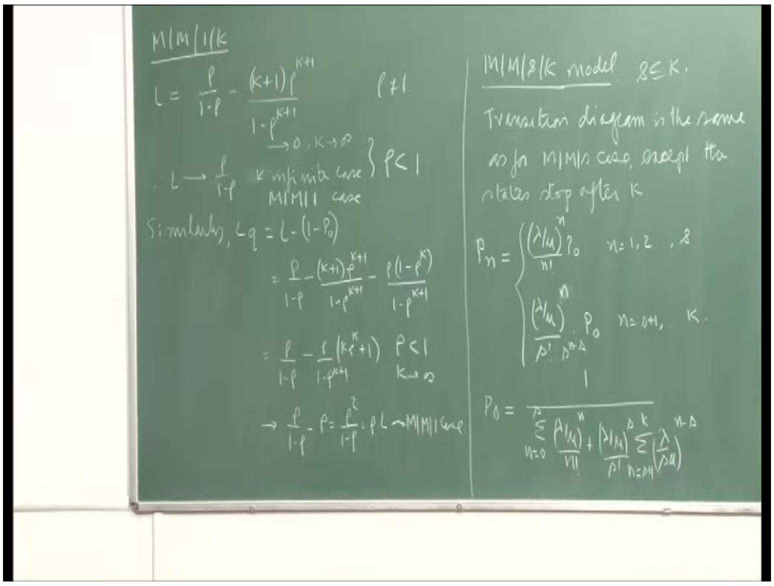
Introduction to Probability Theory and its Applications
Prof. Prabha Sharma
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 37
M/M/I/K and M/M/S/K Models

(Refer Slide Time: 00:14)



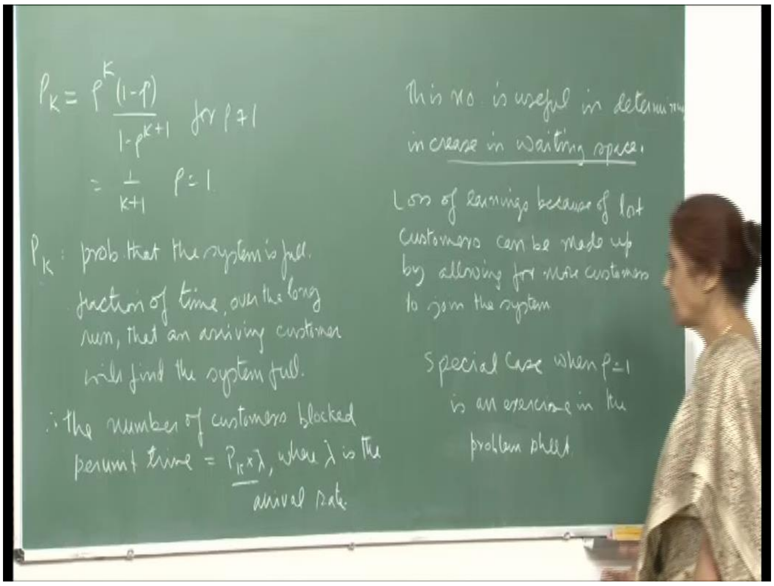
See, the last lecture we talked about m m 1 case, and I will just continue with that and the basic thing to be noted is that see the difference between m m 1 case and m m 1 k case since only the q size. That is you are restricting the number of people in the system. So, for m m 1 k the number of states only go up to k, for m m 1 the states were, it could go up to infinity.

So, therefore, the derivation as I have showed you in the last lecture, the derivation was straight forward. And then see I also showed you that as k goes to infinity, and if of course, here in this case here, we require that o is less than 1, and that I am saying is that the formulae these are all valid, when rho is not equal to 1, because otherwise you cannot divide by 0. So, a for rho less than 1, we showed that as k goes to infinity l will go to rho upon 1 minus rho, which is the k infinite case. That is which is the m m 1 case, as it should be. Because, when you allow your k to become large then it reduces to the m m 1 case, and similarly the derivation for l q. So, this is l q l minus 1 minus p0, you can see that from the formulae for l q because it will be n minus 1 raise to into p n summation. So, you get this.

And therefore, this is the expression, and here again you will see, so I write the expression for l. And, this is for 1 minus p0 and therefore, you can simplify this. And finally as because again this portion will go to for rho less than 1, and as k goes to infinity. So, this portion will become 1. Because, rho is less than 1, so I can again divide by. So, you will be simply left with rho here, k goes to infinity. So, this portion you can show will go to 1.

And therefore, this will be rho upon 1 minus rho, minus rho, which will be equal to rho l. So, again the same as m m 1 case. See, here we can there is so many ways in which you can actually show that this quantity will go to 1, just as we worked out for the l case. The same tricks you can use here, and show that this quantity will go to 1. So, therefore, this is the 1 case. And now just let us look at the probability p k, which is a very important quantity here.

(Refer Slide Time: 03:04)



So, p k is rho k into 1 minus rho upon 1 minus rho raise to k plus 1, when rho is not equal to 1 and it is 1 upon k plus 1 and rho is equal to 1. So, this part of course, I have asked you to do it as an exercise. Now, p k is the probability that the system is full because, you cannot have more than k people in the system. So, p k is the probability that the system is full, which can also be interpreted as fraction of time over the long run.

See remember, we are always talking in terms of steady state probabilities. So, over the long run this also can be interpreted as the fraction of time, that an arriving customer will find the system full because, it already has k people. So, therefore, no more entries can be allowed. And, so this will be a fraction of a time of the time when an arriving customer will find the system full

So, therefore, the number of customers blocked per unit time would be the arrival rate lambda into the fraction of time the system is full. And, so p k into lambda is the is the important number that tells you, the the number of customers blocked by unit time. And, so this is your lost business, lost revenue. So, this is this number is useful in determining increase in waiting space. So, I have been talking about this, what I am saying is that; see now, you can the loss of earnings because, of lost customers can be made up by allowing for more customers to join the system.

So, therefore, now the management has a very important guiding tool, this is you know you can find out that if you are losing that many customers per unit time. Then over day or a week or over a year, whatever your planning horizon, how many customers would have been turned away. And therefore, you can compute estimate the revenue that you would have earned if those customers, who are allowed in and then you can you know, compare it with the cost of increasing your services or allowing for more waiting space and so on.

So, therefore, one can then very comfortably come up to a decision, has to if you lost customers this number is large. Then you would certainly consider increasing your waiting space, and allowing for more people to be serviced, and come to the system. So, as I have been saying that you know these are a really useful models, and they help you. And, the management or the people concern people can always use these as guidelines, not again go by exact number, but at least they can be very good guidelines.

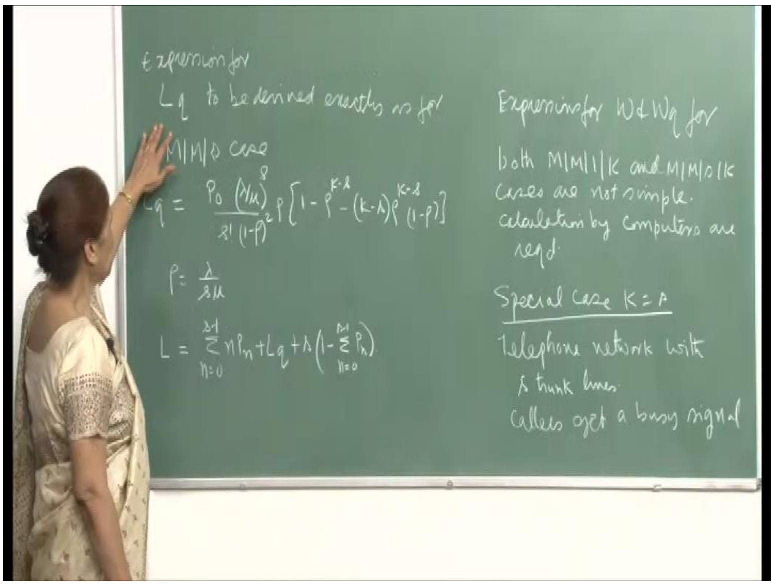
They give very good guidelines. Now, as I said special case when rho is equal to 1 is an exercise, in the problem sheet, which we will be discussing at the end of this lecture exercise nine. So, I have put this you know just for you to compute very simple, but you can then see that the all the quantities can be computed, when rho is equal to 1. That means that is this implies that lambda is equal to mu. So, the service rate, and so here again, there is no question the system growing up because, it is a finite space or a finite space model.

So, now, the case that again I will not go in to detail, but it needs to be mentioned. And, I have written down the formulae for completion sake. So, this is m m s k model. So, now, you have s server s greater than 1, and again finite space. If I not allow more than k people in the system, and certainly your number of servers have to be less than or equal to k, otherwise; it does not make sense. If you are allowing for only ten people to be in the system, and you have 12 servers.

So, certainly the 2 servers will always be idle. So, therefore, s is less than or equal to k and this is nothing really to spend time on arriving all these formulae. Because, the transition diagram is the same as for m m s case, just as here the things were exactly the same as for m m 1 case, except that you had. We have to stop at after state k and here, m m s case these 2 are similar, except that here again you are chopping off after the state k is reached. And, so therefore, the same transition diagram and you can write the same balance equations, and then we can obtain the value for p n.

So, this will be lambda by mu raised n upon n factorial p0, for all values of n from 1 to s, and for values of n greater than s, s plus 1 to k this will be lambda by mu raise to n upon s factorial. Then, 1 upon s raise to n minus s and p0.So, then p0 can be written as this by summing up all the probabilities. So, this is your value for p0. Now, again we will write down the formula for l and l q. The derivation is exactly the same as for m m s, and then you know go on with the other this thing.

(Refer Slide Time: 08:40)



So, when the expression for l q will be derived exactly as for the m m s case, but k is infinity. And, the argument will be the same exactly, you will be summing up sigma n minus s into p n upon n varying from 0 to k, instead of 0 to infinity that is all. So, therefore, that is why this portion has come in.

So, otherwise when k goes to infinity, and for rho less than 1, you will see that this expression will be gone, And, so you will be left with this, which is exactly equal to your l q for m m s case. So, that you can derive that and this is surely your rho is this. And, so far k becoming very large, the your rho has to be less than 1, this has to be less than 1.

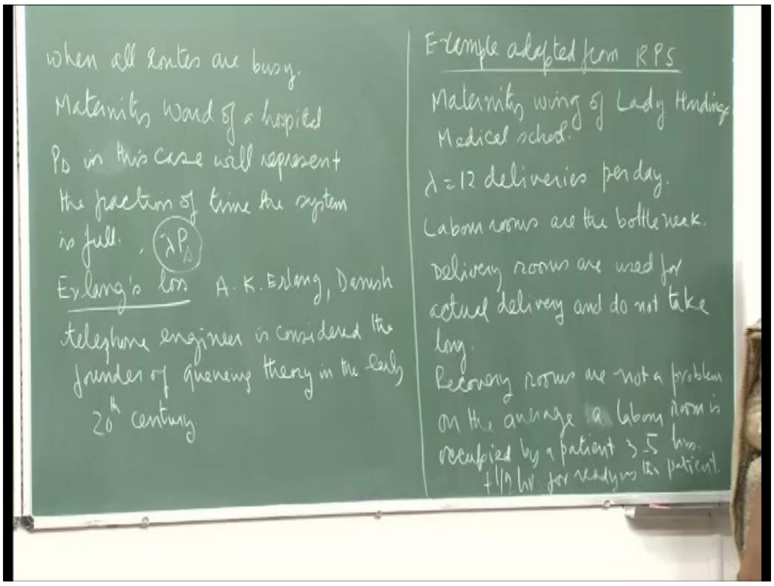
And then the expression for l can be then obtained from l q, and so that is all. I mean you just write down the expressions, I am doing this for the completeness sake. So, that you can yourself if you want to derive them then you can check that the answers are ok. Now, expressions for w and w q for both m m 1 k and m m s k, are not simple. So, what we basically do is; we use the computer calculations to given the values of k, and then value of s and lambda mu. We will simply put these values in, and write a small program to compute the values of w and w q.

And of course, some people in some textbooks you may find, just as we plotted graph for a m m s m m s case. You know we could plot p naught against the values lambda and s and mu, and so on. So, you will find these values tabulated somewhere, now the interesting case out of this is the 1 in which your number of servers is equal to the number of people allowed in the system; that means, you allow only as many people in the system as you have servers. And of course, an immediate example is your telephone network with s trunk lines right.

So, if you have s trunk lines, callers get a busy signal when all routes are busy. So, therefore; obviously, they cannot make a call unless some route becomes free. So, therefore, the special case is of lot of interest, when the number severs is equal to the number of people allowed in the system. And of course, the other extreme case could be, when you know when you go to a restaurant it is self service.

So, then; that means, the number of servers is equal to the number of people in the system. So, that of course, is also there. Then, maternity ward of your hospital because there also see, the number of beds that is available only you can only entertain that many patients, that many women, who are going to deliver? And so you have to turn away other people.

(Refer Slide Time: 11:42)



This case is also of interest, and here, in this case p s will be the will represent the fraction of time the system is full. So, otherwise in these 2 cases it was p k, but now since s is equal to k. So, p s will represent the fraction of time the system is full, and therefore, again you can find out. So, the number of customers lost per unit time will be then given by lambda p s.

If the lambda is your arrival rate then it will be the number of people who are turned away. Now, this is also called the Erlang's loss see, Erlang's, A.K. Erlang was a Danish telephone engineer. And, he is considered to be the founder of queuing theory in the early 20^{th} century. So, you see it is amazing, how as a telephone engineer? He had you know excess to these you know the queuing systems. And therefore, he sort of initiated lot of ideas you know, where these the whole theory has now been developed.

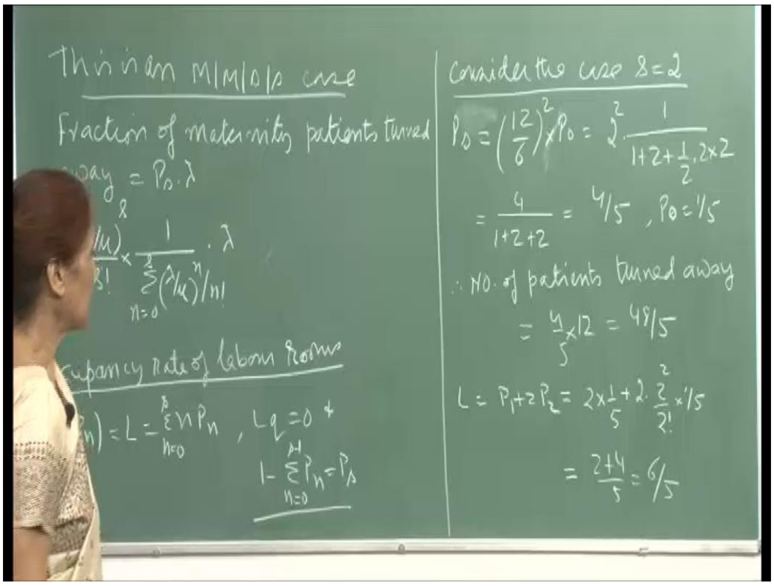
Now, I like to just look at these m m s k case, through an example, and I have just adapted. See, this is from Ravindran, Phillips and Soldberg, but I have adapted it to our system. And so I have just named it the hospital as a lady Harding medical college, which is it is not school, college.

The Lady Harding medical college, which is in New Delhi. And, this is the maternity being of the lady Harding medical college. So, suppose there are 12 deliveries per day. Now, the thing is that only the labour rooms, where the people where there were people come? When the pregnant ladies come with labour pain. So, then they have to wait and then the actual. So, the labour rooms are the bottle neck because, that is where the patient lie till, the delivery has to be has to take place.

So, delivery rooms are used for actual delivery, and do not take long, and the recovery rooms are not a problem. Because, once that a babies delivered then they can be put in a general ward. Whatever; any because there is no special medical attention is needed at that time. So, the recovery rooms are also not a problem.

On the average a labour room is occupied by a patient from 3 to 5 hours, and then half an hour is taken for rating the patient for the delivery. So, therefore, this is where the bottle neck is because, this occupy the longest by a patient. So, and therefore, the idea here is that you can say that may be 6 deliveries per day. So, the deliveries that can. So, 1 labour room can accommodate 6 deliveries, 1 labour room can accommodate 6 patients a day. And of course, your arrival rate is 12 per day.

(Refer Slide Time: 14:58)



So, we just now look at the parameters l q, and the number of patients, who get turned away and so on. So, since the number of labour rooms is the bottleneck. So, we will call the servers, number of servers is the number of a labour rooms. So, therefore, as we saw that system can be modelled as a m m s s.

The maternity being of the lady Harding medical college, which is also has the hospital attached to it. So, that will be treated as an m m s s case since, we are taking the arrival and the service pattern to the morkovain. So, and then the fraction of maternity patients turned away is p s into lambda. So, this is the fraction of time that the system is blocked, and then a lambda is the arrival rate. So, this is the fraction of maternity patients that will be turned away.

Rate of the maternity patient turned away, which the formula for that would be lambda by mu raise to s upon s factorial into 1 upon. So, your p0 would be just adding up to, this whole thing the formula is, we have obtained it number of times by now. So, p0 will be 1 upon this only, because you will not go beyond s. And therefore, this is lambda by mu raise to n divided by n factorial n varying from 0 to s. So, this is the formula for the number of fraction of maternity patients, that will be turned away because, your labour rooms are all occupied.

So, compute the occupancy rate of the labour rooms, we compute expected value of p n, which is for which our notation is l; that means, the average number of people in the system, and this is will be summation j varying from 0 to s j into p j, because either you do not have any patient or you have 1 2 3 up to s only. Because, you only permit as many patients as there are labour rooms. So, therefore, l q is 0 there will no waiting queue, there will be no people waiting in the queue, and summation. So, 1 minus sigma p j j varying from 0 to s minus 1 is equal to p s.

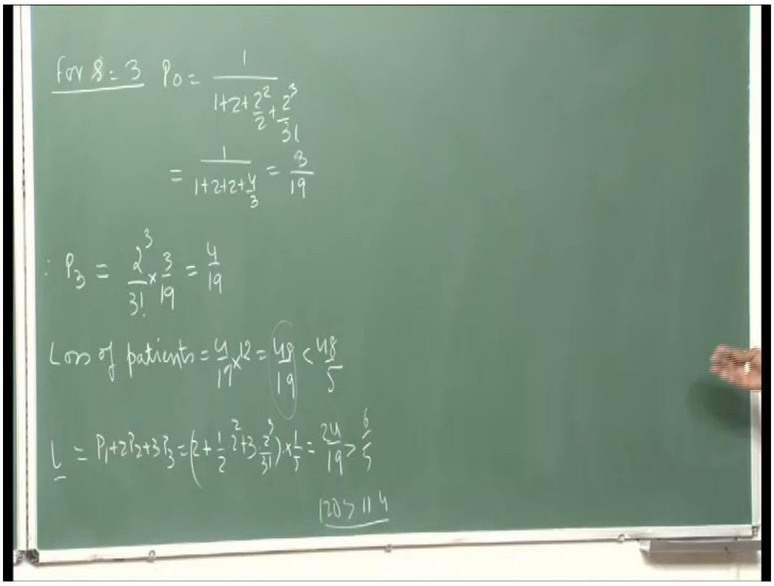
So, with these conditions we can compute this, and that will give us the occupancy rate of the labour rooms. Now, let us compare the ah situation; that means, the hospital has the college the college authorities have a choice, whether to continue with the current 2 labour rooms or to increase the labour rooms. So, that you do not call this comfort, and you do not turn away too many people because, that certainly has a reputation on the hospital if you are saying that all the time your beds are full.

Your labour rooms are busy. So, when s is 2 because your lambda is 12 and your able to the labour room can be used, occupied the number of times it can be occupied is 6. So, we will say that the service rate service rate is 6 and the arrival rate is 12. So, twelve by 6 square into p0 right your p s from here and your p0 will be...

So, this is 2 square upon 1 plus 2 plus 1 by 2 into 2 square, because your s is two. So, just apply this formula, and when you do the computations it comes out to be 4 by 5. Because, your p0 is 1 by 5 and this is 4. So, 4 by 5 is your, which is you can see is high. Four; that means, the fraction of time you turn away patients is 4 by 5. And, the number of patients turned away. So, the rate of number of patients turned away per unit time will be 4 by 5 into 12, which is 48 by 5. I mean per day you are turning away 48 by 5 patients, you are not able to accommodate them.

When you have 2 labour rooms, and these of course, see here this I do not consider this is really important. Because, this will depend on the since your only allowing as many people as there are servers, the number of labour rooms. So, therefore, this is just as you that as you are calling the occupancy rate of the labour rooms.

(Refer Slide Time: 19:31)



This is 6 by 5 now, for s equal to 3, when you make the computations your p0 is comes out to be 3 by 19. Because, now you will go up to 3 the summation, sigma n varying from 0 to 3. So, this is the computation and therefore, it comes out to be, which is 3 by 19. So, therefore, p0 is of course, the probability that there is no there is no patient in the labour rooms, and here your p0 was 1 by 5.

This was 1 by 5, and this is 3 by 19. Then, p 3 comes out to be 4 by 19 and therefore, the loss of patients is 4 by 19 into 12 which is 48 by 19. So, this is definitely less than 48 by 5. In fact, much less than this number. So, therefore, the loss of patients drastically comes down, by increasing 1 labour room. So, with 3 labour rooms, the average number of patients present in the hospital in the maternity ward l, will be given by p 1 plus 2 p 2 plus 3 p 3.

And. So, that will be this number into 3 by 19, which is your p 0 and. So, this turns out to be 30 by 19, which is much higher than 6 by 5, because 150 and this is 114. So, therefore, what is being said is that; which of course, is obvious in the sense that l is higher for 3 labour rooms, then when you had 2 labour rooms. So, this was your traffic intensity or utility whatever you want to call it, for with 2 labour rooms and this is. So, why? Because, you are turning away less patients, and if good will counts then certainly it is more important to have lot of good will in the community, and so 3 labour rooms may be worthwhile, then 2 labour rooms. Know considering the expense of having another labour room.

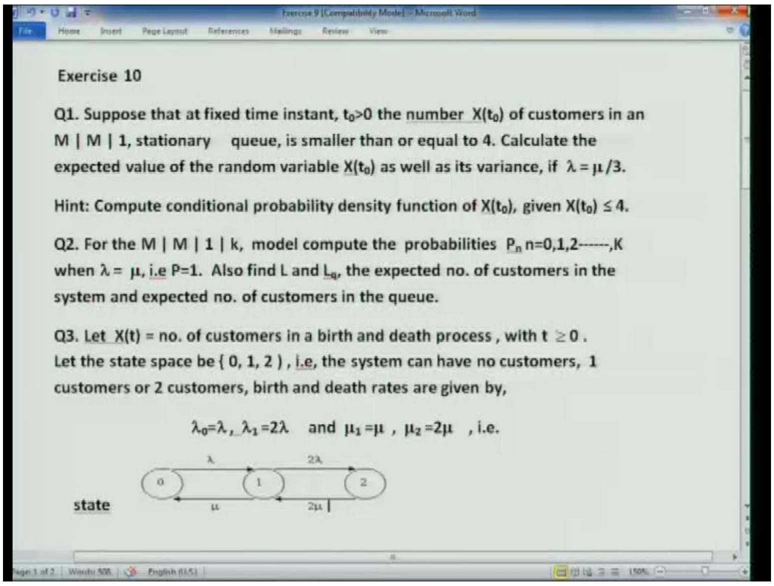
One more doctor 1 more labour nurse and so on, but anyway. So, this is something for the organization to consider, but anyway the numbers tell you something, and this is could measure to see that your utility would be higher 3 labour rooms obviously, but important thing is that you are turning away less patients.

So, that you know the whole idea it was actually in this course, the idea was not to discuss queuing theory extensively. But, I essentially wanted to show you because, after having developed probability theory, I thought it was important if you, get insight into why this theory is. So, useful and therefore, I started talking about you know its applications in the sense that. So, you have seen that throughout your discussion of queuing theory, we have almost used all the concepts of probability theory that we developed. And, other stochastic process is also, that we have the Markova process that we have discussed already.

There also you see that you will be using, that we have used the concepts of probability theory. So, the whole idea was that while you know seeing these applications, you get a good insight, and good understanding of the probability theory. So, that was a basic idea it is not that we would trying to really cover the Markova processes, and queuing theory extensively.

So, now I will just discuss exercise 9 where I have collected some problems, and then we will continue with some more discussions of some more applications of the probability theory we have used, may be through reliability. I want to show you the applications of probability theory to reliability which will come later.

(Refer Slide Time: 23:36)



So, exercise suppose, that at fixed time instance t 0, which is positive the number x t 0 of. So, now, since I have collected these problems, from different books. So, the notation may be a little different. So, here x t0 is the number of customers in an m m 1 stationary queue, we have been refereeing to the number of customers in queue as n t 0. So, does not matter.

And is smaller than or equal to 4. So, at a fixed time the number of customers in an m m 1 stationary queue, is smaller than or equal to 4. Calculate the expected value of the random variable x t0 as well as its variance if lambda is mu by 3. So, you have given rho, rho is 1 by 3, and you are asked to compute the expected value of the random variable. So, therefore, see what are the possible values of x t 0? All you have told is that it is less than or equal to 4. So, therefore, the possible values can be 1, 0, 1, 2, 3 and 4.

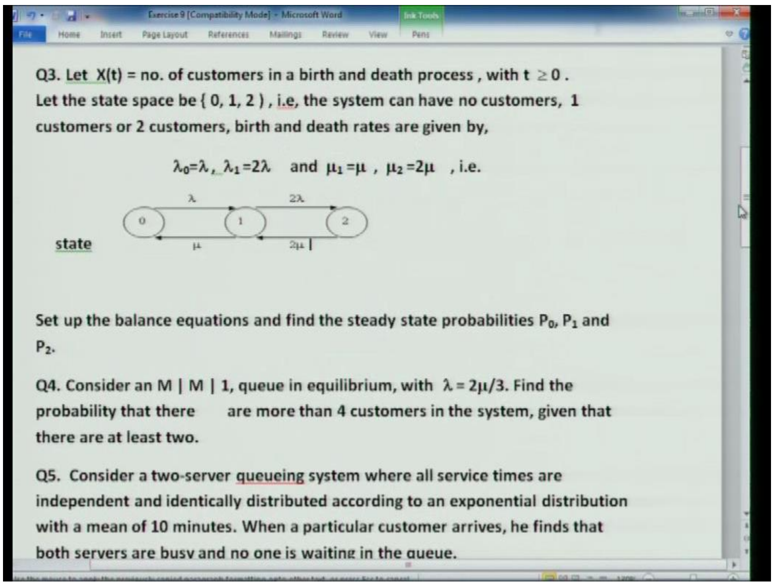
So, we just have to compute the probability, conditional probability, that probability x t 0 equal to say, I given that x t 0 is less than or equal to 4. So, you can do this, whatever we have learnt enough methods to compute your conditional probabilities. So, once you have the conditional p m f of x t 0 you will be able to then find out expected value, and the variance of this random variable.

So, I have given the hint also, I have said that compute the conditional probability density function of x t 0, given that x t 0 is less than or equal to 4. Now, question 2 says that for m m 1 k model, compute the probabilities p n, and 0 1 2 k, when lambda is mu. So, I had said that in the in the lecture also, that when your rho is 1, we want to find out the probabilities, and also the values of l and l q.

So, the value for p and I have already given to you as, 1 upon they all will be the same. So, therefore, it will be 1 upon k plus 1. And now, you have to compute l and l q fine. So, that is straight forward question 3 is that x t 0 is number of customers in a birth and death process, and with t non negative. Let the state space b consisting of 3 states which are 0, 1 and 2 that is the system can have no customers, 1 customer or 2 customers.

So, birth and death rates are given by... So, the transition diagram you can see that lambda 0 will be lambda, but when lambda 1 will be 2 lambda, and mu 1 is mu, mu 2 is mu. So, here they have defined the arrival rate, and the departure rate. So, the arrival rate is the when the 1 then it is 2 lambda it becomes 2 lambda.

(Refer Slide Time: 26:57)

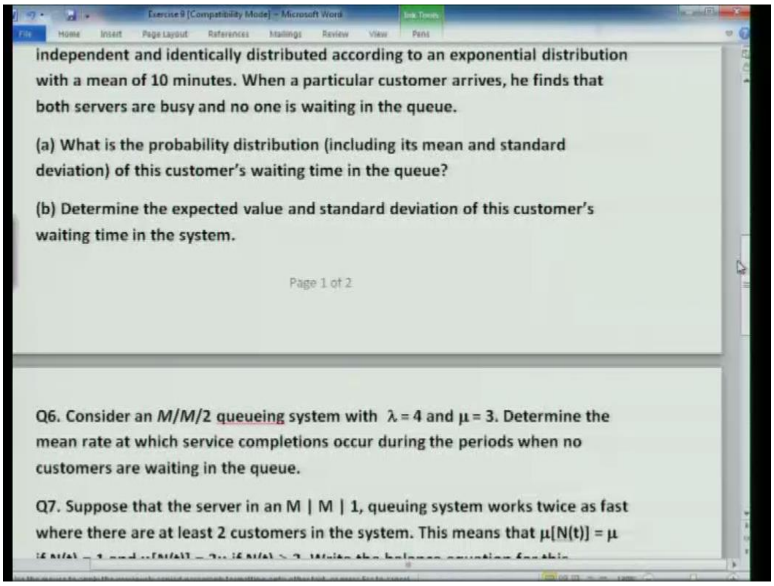


So, set up the balance equations, and find the steady state probabilities p0, p1 and p2. So, I have deliberately given this because, this is departure from your, we have not really discussed the case when lambdas also change. But, since there are only 3 possible states you should be able to write down the balance equations and therefore, then compute your the probabilities p 0, p 1 and p two.

So, I hope you enjoy doing this, then question 4 is again a this is an m m queue in equilibrium, with lambda equal to 2 mu by 3. Find the probability that there are more than 4 customers in the system, given that there are at least two. So, again this is a computation of a this thing so; that means, you are saying that your n t if you know for a fixed time t your saying that n t is greater than or equal to 4, given that n t is...

So, there are more than 4 customers means l t is greater than 4, and then and your given that n t is at least two. So; that means, n t is greater than or equal to two. So, conditional probability of n t more than 4 given that n t is greater than or equal to 2. So, again a simple computation of the conditional probability.

(Refer Slide Time: 28:16)



Now, let us come to question 5, consider a 2 server queuing system. So, they are s s 2, where all service times are independent and identically distributed, according to an exponential distribution with the mean of 10 minutes. Now, the problem 5 and 6, I have taken from hillier and lieberman’s book and. So, therefore, the statements are little different, in the sense that see the service times are independent, and identically distributed, which we have been referring to as markov process. So, with a mean of 10 minutes.

So, remember the mean is 10; that means, the parameter for the exponential distribution will be 1 by 10 when a particular customer arrives, he finds that both servers are busy, and no 1 is waiting in the queue. So, then what is the probability distribution including its mean and standard deviation of this customers waiting time in the queue?

See, you are asked to find out, what is the customer comes in to the comes to the system both the servers are busy? So, you are asked to find the probability distribution of this customers waiting time in the queue. So, w q, when 2 servers are busy, now, what will be the see since the 2 servers are busy; that means, any 1 of them can get serviced, and then his turn will come. So, his waiting time is till, 1 of the service is gets completed right ok.

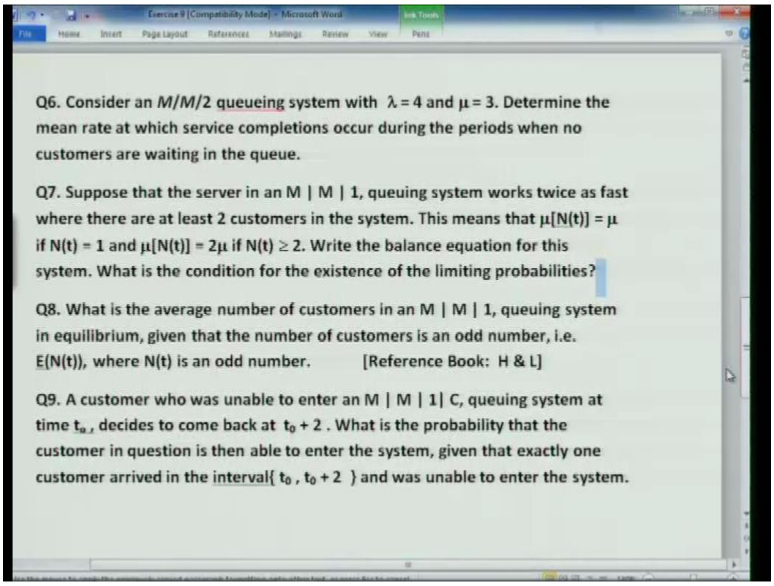
And so therefore, his waiting time is over in the queue the moment 1 of the services is completed. So, now, since there are 2 servers, and each of them is you know with the mean time of 10 minutes. So, once service over; that means. So, see what we have been doing. So, our mu is 1 by 10. So, now, when there are 2 servers our, this thing mu will be 2 by mu 2 by 10. So, which is 1 by 5 and therefore, the corresponding mu will be the mean service time would be 5 minutes.

So, therefore, you can then compute the so; that means, it will again be exponential with mean as 5 minutes. And therefore, you can compute the variance. I am just giving the hints, and within time in the system. Now, you can also determine the expected value, and standard deviation of this customers waiting time in the system. So, in the system when you want to compute that it will be the waiting time plus your mu.

Because, his own service see the mean service time is 10. So, therefore, your w is w q plus mu, and your mu is 10, the mean waiting service time. See, that is not confuse here I am calling mu as 10. So, 1 by mu will be 1 by 10, which is the parameter for the exponential distribution. So, I am calling the mean service time, whichever way you like if you if you still want to refer to mu as 1 by 10, then here it will be 1 by mu. Because, his w q has a mean of 5, and his own service time has an average of 10 minutes. So, then it will be 15 minutes.

So; that means, in the system his mean time mean time in the system would be 15. So, correspondingly you will have mean, and the variance, because for the exponential distribution if mu is if 1 by mu is the mean, then 1 by mu square is the variance. So, you can compute accordingly. So, let us go to problem 6.

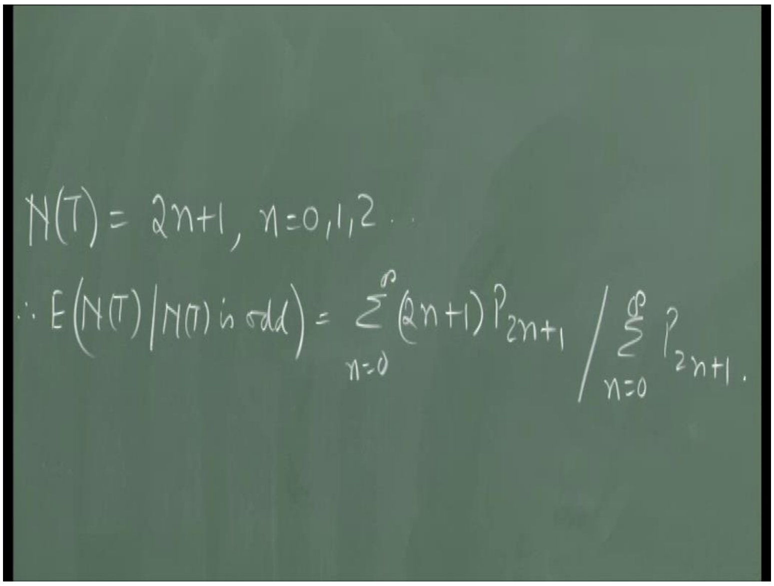
(Refer Slide Time: 32:00)



Now, consider an m m 2 queuing system with lambda equal to 4 and mu equal to 3, determine the mean rate at which service completions occur during the periods, when no customers are waiting in the queue. Determine the mean rate at which service completions occur during the periods, when no customers are waiting in the queue. So, so here it is lambda is 4 mu is 3. And, so you will compute see here, little departure and again I thought, I will give you I have put the problem in to just to show you see you will compute p 0, p 1 and p 2.

Because there are 2 people in the system 2 servers there are 2 servers. So, therefore, either no people in the system 1 person in the system or 2 people in the system. So, mean rate, when no customer in the queue. So, the mean rate, when no customer in the queue I am writing as, the formula for that is... So, mean rate when no customer in the queue.

(Refer Slide Time: 33:09)



So, this I am defining as mu 0 p 0, plus mu 1 p 1 plus mu 2 p 2, which is and divided by p 0 plus p 1 plus p 2. You see, by definition mu 0 is 0 since no service is completed when no 1 is in the system. So, then your mu 1 is 3, but your mu 2 becomes 6 right twice mu mu 1 this is twice mu 1. So, mean rate when no customer in the queue, and this is how we are defining it. So, you compute p 0, p 1 and p 2. And then for an m m 2 queue and you can then write down this.

Now, 7 th is that server in an m m 1 queuing system works twice as fast, where there are at least 2 customers in the system. This means that your mu and t is mu if n t is 1. And so now, I am back to my notation because, here n t is the number of people in the system at time t up to time t. So, therefore, this mu or at time t that is better way to put it. So, mu and t is mu if n t is 1, and mu and t is 2 mu, if n t is greater than or equal to 2. So, here again, just change the system a little, and now, you can write down the balance equations for this system.

And, then what is the condition for the existence of the limiting probabilities, and you can compute the probability. I have kept the thing. So, small and therefore, you can you know do this changes in experimental that. Question a tells, what is the average number of customers in an m m 1 queuing system? In equilibrium given that the number of customers is an odd number, which is you have to find expected value of n t, where n t is an odd number.

So, here again I have taken this problem from hillier and Lieberman, now see remember you have to just compute the expected value. So, your n t will be equal to 2 n plus 1, as n varies from 0 1 to n since we are asking for odd number of people in the system. So, to find the expected value this is the conditional expectation. So, expectation of n t given that n t is odd.

So, which you will write as 2 n plus 1 summation n varying from 0 to infinity, 2 n plus 1 into probability that there are 2 n plus 1 people in the system. And then divided by number of the probability of they are being odd people in the system, which is sigma n varying from 0 to infinity, probability 2 n plus 1.

There is a customer, who was unable to enter an m m 1 c. So, here again c means your k; that means, finite capacity queuing system at time t 0 decides to comeback at t 0 plus 2, what is the probability that the customer in question is then able to enter the system given that exactly 1 customer arrived in the interval, and was unable to enter the system. So, this I will leave for you people to you know really think about it because, this is interesting and it is challenging problem.

So, let us see that you are able to crack it. So, the whole idea is that between t 0 and t 0 plus 2 somebody arrives the system is still full. So; that means, exactly at t 0 plus 2 there is 1 vacancy; that means, one of the services has been completed. And therefore, this person is able to so; that means, at t 0 the system was full and then exactly at t 0 plus 2 the system is empty has 1 vacancy and. So, this person can enter. So, you have to work out this problem. So, I hope you enjoyed doing this assignment sheet.